MATH4100

Li, Yihao Le, Wesley Voss, James Almonte, Marcel

January 2021

1 Introduction

IntelyCare uses data and technology to help long-term care facilities find the extra help they need on a per-shift basis from our pool of credentialed nursing professionals. We give nursing professionals full flexibility to work as much or as little as they choose, and we observe substantial variance in usage across our pool of nurses. Our goal is to identify cohorts of nurses with similar in-app behaviors. Who are nurses that accept whatever shift we show them? Who are the nurses who only work when they find the exact shift they like? Which nurses can work more shifts in a week and which nurses cannot? Who are the weekend warriors? Who are the holiday heroes?

These behavioral cohorts will allow us to personalize the nurse experience in important ways. We can send more timely and relevant communications, improve our in-app shift recommendations, and offer more effective promotions if we understand what behavior patterns characterize a given nursing professional.

To achieve this goal, we will use some unsupervised learning methods and supervised learning techniques form machine learning filed.

2 Methodologies

2.1 Unsupervised learning

Unsupervised Learning is a method of machine learning in which no pre-labeled training examples are given, for input data is automatically classified or grouped. The main applications of unsupervised learning include: cluster analysis, relationship rules, dimensionality reduction etc. This week, we studied some techniques or methods of unsupervised learning. The first one is principal component analysis (PCA). PCA is an exploratory tool to simplify a large and complex data set into a smaller, more easily understandable data set. It summarizes the complex data set by creating variables that are linear combinations of the original data set. It does not necessarily describe or explain the relationship among variables. After constructing principal component analysis, each of new variables, called principal components, is uncorrelated with all others. It is effectively used as a method of pattern analysis bringing strong patterns to the forefront.

In summary, PCA looks to find a low-dimensional representation of the observation that explain a good fraction of the variance. The other technique is clustering. There are two clustering techniques discussed in this book. The first one is k-mean clustering. This approach is aimed to seek to partition the observation into a pre-specified number of clusters. Despite the complexity of k-mean clustering, there is a simple algorithm addressed in the book to transfer such a massive question into an optimization one, which is finding the closest Euclidean distance. However, in order to perform the k-mean clustering, we have to determine the number of clusters we expect in the data. In other words, there is a huge practical consideration that arises in this case. On the other hand, the aim of the hierarchical clustering is similar but it does not require the particular choice of K right in the beginning.

2.1.1 Clustering Methods

Depending on the clustering method that is used, there are certain criterias that determine the output.K-Means and spectral clustering is a method of clustering where the user is able to determine the number of desired clusters. There are many others, but this is getting long. Reference the clustering resource for the full list.

On the surface, data can be supervised and unsupervised. When dealing with supervised data the goal is to infer from the training set by means of regression and classification. These regressions can be linear or polynomial. In terms of classification one can classify data by means of logistic regression, k nearest neighbors, and decision trees. When it comes to regression the purpose is to predict numerical values, yet in terms of classification the purpose is to predict categorical values.

2.2 Data Visualization

For data visualization, we used the following two methods.

2.2.1 T-distributed stochastic neighbor embedding

T-distributed stochastic neighbor embedding (TSNE), is a model used to put data with many dimensions into a 2D, or 3D plane. This happens in two steps, first "High dimension Probability distribution" is necessary. This means that data points that are similar, have a lower probability and will be physically closer together, creating a higher density of points. The data that is not similar, will be physically further away from others, with a high probability. High and low density is an effective metric when it comes to clustering. By viewing the data in multiple dimensions it becomes easier to assign a probability distribution to the data. When reducing dimensions, it is crucial to be very familiar with how the different parameters affect the data. These parameters have a direct effect on the outcome, meaning that if you are not familiar enough with your technique, you could draw false conclusions and have no idea. Additionally by

over classifying or overtraining the adaptability and plasticity of the model when it comes to receiving new inputs can be reduced.

t-SNE computes probability p_{ij}:

$$p_{i|j} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2)}$$

"As Van der Maaten and Hinton explained: 'The similarity of datapoint x_j to datapoint xiis the conditional probability, $p_{i/j}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i."

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

and note that $p_{ij} = p_{ji}, p_{ii} = 0,$ and $\sum_{i,j} p_{ij} = 1$

2.2.2 Multidimensional Scaling

Multidimensional Scaling (MDS) is a way to present data to show similarity between individual cases within a data set. It creates new variables that still represent the original variables, as well as make it so that the data can be represented in a 2-D display with N dimensions, highlighting the difference, or distance, between the points.